MatInspector

The basic methodology of matInspector has been reported (Quandt et al., 1995; Kel et al, 2001; Cartharius et al., 2005). Briefly, for a transcription factor with a set of known binding motif sequences, a PWM may be constructed by counting the numbers of bases, A, C, G and T, at each position. After normalizing the counts, the resulting PWM, designated as ***F*** (see an example below), is a $4 \times w$ matrix with elements $f_{x,i}$, where *w* is the length of the motif.

	1	2	.	i	.	w
A	$f_{1,1}$	$f_{1,2}$	.	$f_{1,i}$	.	$f_{1,w}$
C	$f_{2,1}$	$f_{2,2}$	.	$f_{2,i}$	.	$f_{2,w}$
G	$f_{3,1}$	$f_{3,2}$	.	$f_{3,i}$	.	$f_{3,w}$
T	$f_{4,1}$	$f_{4,2}$	.	$f_{4,i}$	.	$f_{4,w}$

The PWM is used to scan new sequences for putative binding motifs using a sliding window of length *w* as described below. For each segment of sequence in the window, $m = (x_1, x_{2,} \cdots, x_w)$, x_i is the *i*th observed base pair in segment *m*. To simply the notation, let *x* take the value of 1, 2, 3 or 4 corresponding to base pair A, C, G, or T. The score of the segment is calculated as follows,

$$s(m) = s(x_1, x_{2,} \cdots, x_w) = \frac{\sum_{i=1}^{i=w} I(i) \cdot f_{x_i,i} - \sum_{i=1}^{i=w} I(i) \cdot f_i^{\min}}{\sum_{i=1}^{i=w} I(i) \cdot f_i^{\max} - \sum_{i=1}^{i=w} I(i) \cdot f_i^{\min}}, \quad (1)$$

where $f_i^{\min} = \min(f_{x,i})$ and $f_i^{\max} = \max(f_{x,i})$, are the minimal and maximal base frequencies at position *i* in the $4 \times w$ matrix (*i*th column's *min* and *max*), respectively. $I(i)$ is the information for position *i*, defined as follows,

$$I(i) = 1/\log 10 + (\sum_{x_i=1}^{4} f_{x_i,i} \cdot \log(f_{x_i,i}) + \log 10), \quad (2)$$

For a sequence of length *L*, the algorithm computes $s(m)$ for each of the $L - w + 1$ sites by moving a sliding window of length of *w* through the sequence for both the plus and minus (reverse complementary) strands of the sequence. If a motif scores higher than or equal to a pre-defined cutoff, a putative binding site is declared.

PWMs for Oct4/Sox2

In six well-characterized mouse Oct4/Sox2 target genes (*Fgf4*, *Utf1*, *Oct4*, *Sox2*, *Fbx15*, and *Nanog*), the Oct4 and Sox2 binding sites are separated by 0-3 bases (for a review, see

Niwa 2001). We created two PWMs, one for Oct4 and one for Sox2 using eight binding site motifs. The lengths of the PWMs for Oct4 and Sox2 are 8 and 7, respectively.

References

Cartharius,K. *et al*. (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, **21**, 2933-2942.

Kel,A.E. *et al*. (2001) Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors. *J. Mol. Biol*., **309**, 99-120.

Niwa,H. (2001) Molecular mechanism to maintain stem cell renewal of ES cells. *Cell Struct. Funct*., **26**, 137-148.

Quandt,K. *et al*. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res*., **23**, 4878-4884.